LAION-C: AN OUT-OF-DISTRIBUTION BENCHMARK FOR WEB-SCALE VISION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) robustness is a desired property of computer vision models. Improving model robustness requires high-quality signals from robustness benchmarks to quantify progress. While various benchmark datasets such as ImageNet-C were proposed in the ImageNet era, most ImageNet-C corruption types are no longer OOD relative to today's large datasets scraped from the web, which already contain common corruptions such as blur or JPEG compression artifacts. Consequently, these standard benchmarks are no longer well-suited for evaluating OOD robustness in the era of web-scale datasets. Indeed, recent models show saturating scores on ImageNet-era OOD benchmarks, indicating that it is unclear whether models trained on web-scale datasets truly become better at OOD generalization or whether they have simply been exposed to the test distortions during training. To address this, we here introduce LAION-C as a benchmark alternative for ImageNet-C. LAION-C consists of six novel distortion types across five severity levels designed to be OOD, even for web-scale datasets such as LAION. In a comprehensive evaluation of state-of-the-art models, we find that the LAION-C dataset poses significant challenges to contemporary models. We additionally conducted a psychophysical experiment to evaluate the difficulty of our proposed corruptions for human observers, enabling a comparison of models to lab-quality human robustness data. We observe a paradigm shift in OOD generalization: from humans outperforming models, to the best models now matching or outperforming the best human observers.

1 INTRODUCTION

Vision models have been a cornerstone of modern machine learning, driving breakthroughs in diverse applications. In recent years, large-scale vision models such as vision transformers (Dosovitskiy et al., 2021) and ConvNeXt (Liu et al., 2022), trained on expansive web-scale datasets like LAION (Schuhmann et al., 2022), have pushed the boundaries of performance on standard benchmarks. However, the continued advancement and reliable evaluation of these models depends on the availability of datasets that effectively challenge model robustness and generalization capabilities.

ImageNet-C (Hendrycks & Dietterich, 2019) has long stood as the de facto standard for OOD evaluation, particularly for models trained on ImageNet (Russakovsky et al., 2015). It contains images that are systematically different from those in ImageNet, meaning that models trained on ImageNet must robustly generalize to perform well on ImageNet-C. Previous work (e.g., Hendrycks & Dietterich, 2019) found that OOD generalization is not trivial to achieve: Many vision models do indeed struggle with OOD datasets like ImageNet-C even if they perform well on ImageNet. Hence, these types of unfamiliar inputs are crucial for evaluating the robustness of machine learning models since they are indicative of performance on unexpected input; a challenge that many deployed models face. Modern models trained on much larger web-scale datasets, e.g., CLIP (Radford et al., 2021), exhibit much better performance on classic OOD datasets than IN-trained models, potentially suggesting that they have learned robust representations which better generalize to unseen data.

However, as modern training datasets are scaled well beyond ImageNet, existing OOD benchmarks might not be truly OOD with respect to web-scale datasets anymore. OOD datasets such as ImageNet-C were explicitly created to be OOD with respect to the most popular dataset at that time: ImageNet. ImageNet-C contains images with corruptions potentially relevant for practical ap-

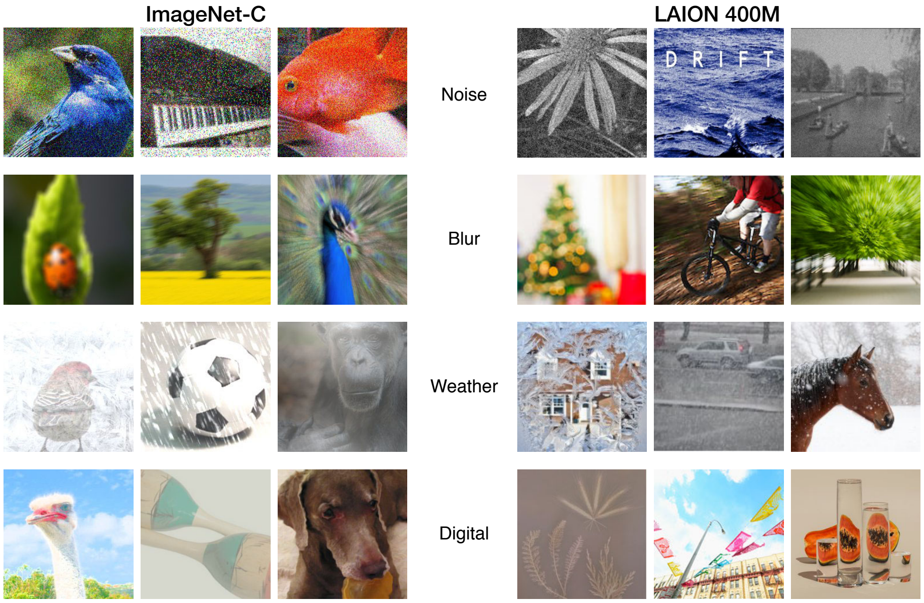

Figure 1: **ImageNet-C corruptions are not out-of-distribution (OOD) for web-scale datasets like LAION-400M.** Exemplary corrupted images from ImageNet-C (left) are similar to LAION-400M samples (right). Each row shows example corruptions and dataset images for one ImageNet-C corruption category (Noise, Blur, Weather, Digital). The presence of these distortions in web-scale datasets indicates the need for an OOD benchmark in the era of web-scale vision models.

plications but (by design) not contained in ImageNet and, thus, OOD. However, with the change of the reference dataset from ImageNet to web-scale datasets such as LAION, these corruptions might no longer be OOD. For example, many images in LAION are blurry—not by deliberate design, but because LAION images were not sampled from a few websites with (implicit) quality standards, like ImageNet images were, but from almost any publicly accessible website online. Simply put, models trained on LAION might have seen the types of corruption on which they are tested with ImageNet-C. For another type of OOD benchmark, namely distribution shifts defined by the style of an image, recent work empirically shows that such datasets are not OOD but overlap with LAION-400M (Mayilvahanan et al., 2023; 2024). This raises a central question: Are modern vision models genuinely improving in terms of OOD generalization, or are they simply trained on datasets that already contain the corruptions, essentially testing in-distribution rather than OOD generalization? This distinction is crucial because if these modern models were not truly more robust than standard models, they might also not perform better on the real OOD test data one might face in practice.

Given the importance of OOD generalization in practice, there is a pressing need for a new benchmark that more effectively evaluates the OOD robustness of state-of-the-art foundation models: an OOD dataset for the era of web-scale vision models. Our **contributions** are as follows:

1. Given that existing OOD datasets are often no longer OOD for models trained on web-scale datasets, we introduce LAION-C, a **novel benchmark** dataset with six manually designed corruption types and 16 superclasses to evaluate the robustness of web-scale vision models.
2. We conduct a comprehensive performance analysis of various models on LAION-C and report a robust human OOD generalization baseline obtained through **psychophysical experiments** with 19 participants, collecting 11,400 trials in a highly controlled laboratory environment.
3. The resulting data serves as an OOD benchmark for current and future models, enabling not only an assessment of their generalization ability on truly OOD data but also providing insights into the **discrepancies between human and machine perception**, observing a paradigm shift in OOD generalization: from humans outperforming models to the best models now matching or outperforming the best human observers.

A detailed related work section can be found in Appx. A.1.

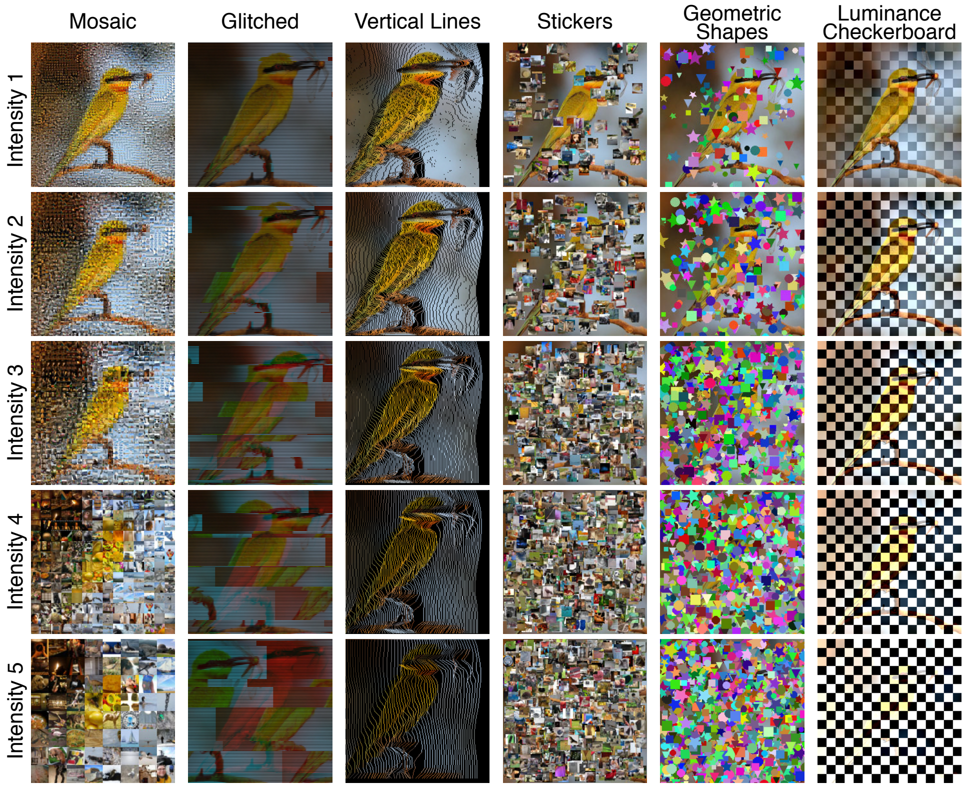

Figure 2: **LAION-C distortions, intended to be OOD even for web-scale datasets.** This figure illustrates the six LAION-C distortions at five intensity levels. Best viewed on screen.

2 METHODS

2.1 CONSTRUCTING NEW OOD DISTORTIONS

As described in the introduction and depicted in Fig. 1, ImageNet-C is not OOD for models trained on large-scale datasets. Given the limitations of existing benchmarks like ImageNet-C, we develop a novel dataset specifically designed to challenge these foundation models more rigorously. Our dataset introduces six carefully designed, fully synthetic distortions that models have not encountered during training. These distortions are designed to be OOD even for web-scale datasets (as supported by quantitative evidence presented later). Hence, models truly need to generalize beyond their training distributions to perform well on this benchmark which we term LAION-C.

Distortions The core idea behind our distortions is to intentionally disrupt visual consistency and perceptual cues that models rely on for image classification, such as texture (Geirhos et al., 2019). Following ImageNet-C, each distortion consists of five different *intensity levels*. The distortions capture a range of visual challenges ranging from disruptions of local image information to more global structural alterations, as described below and illustrated in Fig. 2.

- **Mosaic:** The original image is broken down into smaller tiles, each replaced by a chromatically similar picture. This patchwork creates a mosaic effect that disrupts edges and textures while introducing contextually irrelevant information.
- **Glitched:** The original image undergoes an artistic digital corruption with horizontal lines overlaying shifted image segments and color channel shifts. This dislocates the global contextual structure of the image. While the concept of such glitchy images has been explored in earlier work (Kaufmann et al., 2019), our transformation introduces a more intense corruption.

- **Vertical Lines:** The original image is deconstructed into bent vertical line segments. This distortion retains the original colors but strips away local information, disrupting the contours and edges of the image and introducing visual discontinuity.
- **Geometric Shapes:** The original image is overlaid with overlapping geometric figures such as squares, circles, and stars. This visual clutter introduces local noise that obscures the main object, like the Kaleidoscope corruption from Kaufmann et al. (2019).
- **Stickers:** The original image is augmented with assorted image patches. This addition of visual elements masks features of the primary object.
- **Luminance Checkerboard:** The original image is divided into a grid, with the luminance of each cell altered in a checkerboard pattern. The stark luminance contrast between adjacent tiles and artificial grid boundaries makes this distortion challenging.

We intend to build a challenging dataset that has the potential to guide the future development of vision models. Our dataset incorporates these tougher and less common distortions to simulate the difficulty of OOD scenarios that models might encounter in real-world applications. We tune the intensity levels of each distortion such that either humans or a contemporary vision model (ViT-B) achieve chance performance on the highest intensity level, i.e. no model is expected to perform well on the hardest levels. The other intensity levels are chosen so that we can observe a gradual decline in accuracy, ensuring that the distortions are sufficiently challenging.

These distortions are then applied to a carefully curated subset of images from the ImageNet validation dataset. To contextualize model performance, we later want to compare it to human performance. As human evaluations on datasets with hundreds of classes cannot be scaled to sufficiently many participants, we follow previous work (Geirhos et al., 2018) and simplify the task to a 16-class classification problem. We extract 285 ImageNet-classes to form 16 superclasses, namely ball, bird, boat, bottle, butterfly, car & truck, cat, chair, dog, fish, fruit, instrument, primate, snake, timekeeping, and tool. For robust statistical analysis, our dataset comprises 273 images for each superclass. This data size selection allows us to ensure that a 3% difference in the performance between models is statistically significant. Our dataset serves as a proxy for the unforeseen OOD environments future models must handle, advancing the state of robustness evaluation. Additionally, we manually filter the dataset to ensure that none of the images in one superclass contain objects from another class or require specific cultural knowledge for classification. This process helps to avoid ambiguous ground-truth labels.

2.2 MEASURING MODEL PERFORMANCE

We use the generated datasets to evaluate the performance of a suite of 58 vision models. Our selection includes models trained on large-scale web datasets and fine-tuned on ImageNet-1k, such as Vision Transformers (ViT) (Dosovitskiy et al., 2021), ConvNeXt (Liu et al., 2022), and EVA (Fang et al., 2023; 2024). For comparison, we also evaluate the performance of smaller-scale model families such as ResNet (He et al., 2016) and MobileNet (Howard, 2017) and large-scale models trained only on ImageNet-1k. Additionally, we also evaluate GPT-4o (OpenAI, 2024) and Gemini 1.5 Pro (Team et al., 2024) on a representative subset of LAION images. See Tab. 6 for a complete list of all models we evaluate. To address the imbalance caused by distinct numbers of subclasses within each superclass, we compute the average probability values across subclasses for each superclass, a method first suggested by Geirhos et al. (2018). This adjustment mitigates biases introduced by the varying subclass distributions, ensuring a more accurate model performance evaluation.

2.3 COLLECTING HUMAN PERFORMANCE VIA LAB EXPERIMENTS

To explore the discrepancies between human and machine perception, we design a psychophysical experiment to gather human classification data on the augmented images. This experiment builds on previous paradigms (Geirhos et al., 2018; 2021) to ensure consistency and comparability. In the experiment, 19 human subjects are briefly presented with a distorted image and are asked to classify it into one of 16 classes, reminiscent of how a DNN might be evaluated on a classification task (in contrast to e.g. open response paradigms, where participants could give arbitrary textual responses). Participants were recruited from the university student body, and screened for normal or corrected-to-normal vision and absence of color blindness. The experiment was conducted in a controlled dark environment using a 22” VIEWPixx 3D light LCD monitor, with stimuli presented centrally at a fixed viewing distance to ensure foveal presentation. The classification task was divided into two warm-up blocks and ten main experiment blocks, with each block containing images from 16 superclasses. Participants were given 2.5 s to view each image, followed by a 2 s response window to classify the image by clicking on a set of icons. To motivate high performance, a monetary bonus was awarded for surpassing fixed, pre-determined performance thresholds for each block. Further methodological details are provided in Appx. A.2.

2.4 QUANTIFYING HUMAN-MACHINE ALIGNMENT VIA ERROR CONSISTENCY.

To quantify the alignment between human and machine visual perception, we adopt the error consistency metric proposed in Geirhos et al. (2020b). Error consistency, denoted as $\kappa \in [-1, 1]$, measures the degree of agreement between the classification mistakes of two observers. In brief, κ takes on a value of 1.0 if two observers are perfectly consistent, i.e. if they make classification mistakes on exactly the same images. Two independent binomial observers that agree no more than expected by chance will result in a κ of 0, while two maximally inconsistent observers will have a κ of -1. See Appx. A.3 or Geirhos et al. (2020b) for a detailed explanation of the metric.

3 RESULTS

3.1 HOW OOD IS LAION-C?

Now that we have outlined the construction of our LAION-C dataset, we empirically evaluate whether it is indeed OOD relative to the large-scale image datasets used to train modern vision models. Rigorously quantifying how OOD a test dataset is with respect to some training dataset requires a precise definition of the test and training domain (Mayilvahanan et al., 2024). As the distribution shifts introduced by the distortions of our proposed LAION-C and ImageNet-C are fuzzy in nature, we use three tools to compare the OOD-ness of our proposed dataset to the OOD-ness of ImageNet-C. First, we use a qualitative assessment. By searching for the name and related concepts of ImageNet-C corruptions, we easily find LAION samples with visual distortions akin to those of ImageNet-C samples (see Fig. 1).

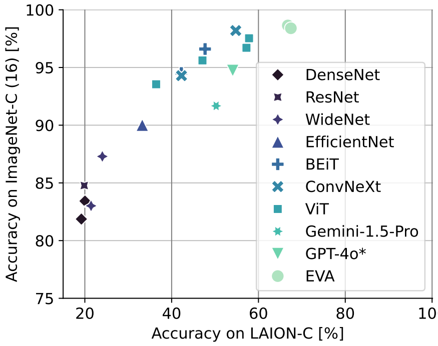

Figure 3: **Performance Divergence of Models on LAION-C and ImageNet-C 16 class.** Evaluating models on the 16-class versions of ImageNet-C and LAION-C produces a plateaued performance on ImageNet-C, while LAION-C still yields a high variance across models.

Second, we use the difficulty of a test dataset (measured by the performance that models trained on a reference dataset yield on the test dataset) as a proxy for how much the test dataset differs from the reference dataset. Here, the reasoning is that if a test dataset can be solved almost perfectly by a model, it means that either the model has great generalization skills or the test dataset is not strictly OOD. If, at the same time, another dataset has much greater difficulty according to the same models, the second dataset is likely more OOD than the first. For the sake of comparability, we here use a version of ImageNet-C restricted to the same 16 superclasses that were used for LAION-C, where we implemented the ImageNet-C augmentations through the code by Michaelis et al. (2019). Indeed, a comparison of the performance achieved by our suite of models (see Fig. 3) suggests that LAION-C is more OOD to LAION than ImageNet-C is.

Third, we use the FID (Heusel et al., 2017; Kynkäänniemi et al., 2022) to quantify the difference between LAION and ImageNet-C and LAION-C, respectively. Specifically, we employ a CLIP-trained ViT-B as feature encoder and use the implementation by Parmar et al. (2022) to calculate FID-scores. In line with the previous evidence, we find a FID of ≈ 70 between LAION and LAION-C, which is substantially higher than that between LAION and ImageNet-C (≈ 40). This means that features of LAION are closer to those of ImageNet-C than those of LAION-C, again highlighting the

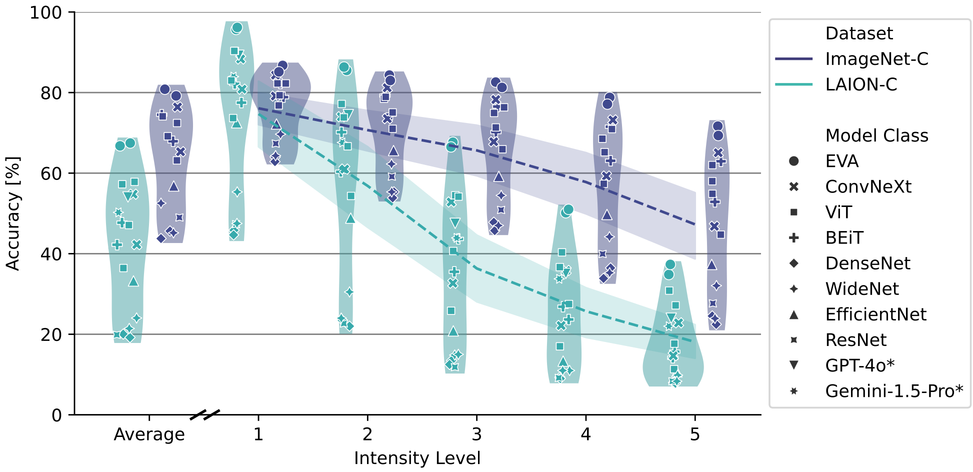

Figure 4: **LAION-C poses a greater challenge to model robustness than ImageNet-C.** We plot distortion intensity against each model's average accuracy. Visual foundation models evaluated on ImageNet-C maintain high accuracy, with minimal drop across increasing intensity levels. On our LAION-C dataset, the models exhibit a sharper decline in accuracy, highlighting the benchmark's effectiveness in measuring model robustness.

greater OOD-ness of LAION-C. In summary, we have presented three different kinds of evidence suggesting that LAION-C is more OOD than ImageNet-C to LAION.

3.2 MACHINE PERFORMANCE

In Fig. 4, we compare model performance on ImageNet-C against performance on LAION-C. Evidently, the average model performance on ImageNet-C stays above or close to 60%, indicating that current models are increasingly adept at handling the distortions in ImageNet-C. This observation reinforces our hypothesis that the challenge presented by ImageNet-C may no longer be sufficiently difficult to rigorously test the robustness of modern models.

In contrast, models achieve much lower accuracy on LAION-C on average and exhibit more inter-model variability. This showcases our dataset's ability to uncover nuances that remain hidden on more saturated benchmarks. These performance differences are particularly obvious at higher intensity levels, illustrating LAION-C's potential to serve as a more challenging and insightful benchmark for evaluating robustness.

We also provide a detailed breakdown of non-averaged, dataset-specific results in Fig. 10. We observe significant variability in the performance of different vision models across various datasets and distortion levels, highlighting the diversity in model robustness. These results further highlight the effectiveness of our datasets in eliciting different responses from models of different architectures, parameter scales, and training data sizes. This diversity is particularly valuable for understanding which model designs are more robust to specific types of distortions, offering insights that are beneficial for advancing the state-of-the-art model robustness.

In Tab. 1, we present a comprehensive evaluation of our suite of models on LAION-C. We report each model's top-1 accuracy on the (undistorted) ImageNet validation set as a baseline (*Clean Accuracy*) and the average top-1 accuracy on LAION-C averaged across distortion types and intensity levels (*LAION-C*). We then break the latter down into the six distortion types. This enables a thorough comparison of model performance, highlighting which architectures generalize best.

3.3 IS LAION-C A PROXY FOR MORE REALISTIC DISTRIBUTION SHIFTS?

To demonstrate that model performance on LAION-C is indicative of real-world performance despite the highly synthetic nature of our corruptions, we analyze the correlations between model

Table 1: **LAION-C benchmark results.** Numbers show the top-1 accuracy in percent. *ImageNet* refers to model accuracy on the (uncorrupted) ImageNet validation set, with values sourced from the timm leaderboard (Wightman, 2024). For each corruption, we report the mean top-1 accuracy across all intensity levels, with *LAION-C* as the overall benchmark metric (averaged across corruption types). GPT-4o and Gemini 1.5 Pro were evaluated on 48,000 images, 100 for each class. For full model names and descriptions, see Tab. 6 in the Appendix.

Model	ImageNet	**LAION-C**	Mosaic	Vertical	Glitched	Luminance	Geometric	Stickers
EVA-G-P14-560-M30M-IN22K	89.8	**67.5**	48.8	53.6	70.8	97.2	**81.0**	**53.4**
EVA02-L-P14-448-MIM-M38M-IN22K	**90.1**	66.8	**53.6**	**58.2**	**78.2**	93.6	76.4	40.6
ViT-H-P14-336-CLIP-LAION-IN12K	88.6	57.3	45.2	51.2	69.8	88.2	64.4	24.6
ViT-L-P14-224-CLIP-OpenAI-IN12K	88.3	57.8	52.6	49.8	68.2	**98.6**	55.4	22.4
ViT-B-P32-384-CLIP-LAION-IN12K	85.4	36.4	36.8	35.2	35.8	54.0	37.6	19.2
ViT-B-P16-224-AugReg-IN21K	85.5	47.1	46.4	42.8	62.0	71.4	42.4	17.6
BEiT-v2-L-P16-224-IN1K	87.4	47.7	52.4	44.8	63.2	70.2	11.8	43.8
BEiT-v2-B-P16-224-IN1K	85.6	42.2	46.2	40.4	52.6	68.2	11.4	34.6
ConvNeXt-XXL-CLIP-LAION-IN1K	88.6	54.8	53.0	53.4	71.8	77.4	52.2	20.8
ConvNeXt-B-CLIP-LAION-AugReg-IN12K	87.6	42.3	37.6	43.8	44.4	54.2	50.4	23.2
WRN101-2-TV-IN1K	78.8	21.4	30.4	28.4	22.0	22.8	18.2	6.8
WRN50-2-RACM-IN1K	81.5	24.0	26.8	21.4	17.0	45.0	24.6	9.4
RN50-A1-IN1K	81.2	19.9	28.0	18.8	20.8	23.4	21.2	7.0
EFF-B3-RA2-IN1K	82.3	33.2	32.4	31.8	40.2	45.2	37.6	12.2
DN201-TV-IN1K	77.3	19.2	28.6	26.2	13.2	23.2	16.8	7.2
DN161-TV-IN1K	77.3	20.0	31.0	26.8	15.2	25.2	15.4	6.6
GPT-4o	-	*54.1*	*42.8*	*45.4*	*65.1*	*80.1*	*54.2*	*36.5*
Gemini 1.5 Pro	-	*50.2*	*34.9*	*37.0*	*46.2*	*84.4*	*60.9*	*38.1*
Best human observer	-	55.2	58.0	55.3	78.7	63.4	40.4	35.7
Average human observer	-	47.0	50.8	43.6	71.0	53.1	34.3	29.1

Table 2: **LAION-C is challenging but can be solved by fine-tuning on the exact distortions.** We compare the performance of ViT-H-P14-336-CLIP-LAION-IN12K before and after fine-tuning it on ImageNet-1k training images with LAION-C corruptions. As the performance after fine-tuning is high, this means that LAION-C, although challenging, remains solvable as it retains enough signal when applying distortions.

Accuracy [%]	Mosaic	Vertical Lines	Glitched	Luminance	Geometric	Stickers
Before	45.2	51.2	69.8	88.2	64.4	24.6
After	79.0	93.5	95.8	97.7	90.2	61.0

accuracy on LAION-C and on several well-established OOD benchmark datasets such as ImageNet-R, ImageNet-A and ImageNet-Sketch in Tab. 5. Clearly, models that achieve high accuracy on LAION-C are also robust to other distribution shifts. However, our main goal is not to measure real-world performance, but to measure a model's ability to generalize beyond its training data, which requires a truly OOD test set - a requirement that might even be incompatible with the requirements of a real-world distribution shift.

3.4 CAN LAION-C BE SOLVED?

Given the low performance of current state-of-the-art models on LAION-C, one might wonder whether LAION-C is simply impossible to solve because the distortions destroy all information necessary for correct classification of the images. To disprove this hypothesis and highlight the validity of LAION-C as a benchmark for evaluating model robustness, we conduct a fine-tuning experiment to assess whether the challenges posed by LAION-C are solvable at all. Specifically, we fine-tune a ViT-Huge model, which was originally pretrained with a CLIP-objective on LAION-2B and then fine-tuned on ImageNet-22k and ImageNet-1k. For this experiment, we use a custom dataset sub-sampled from the ImageNet-1K training set and augmented with the distortions introduced in LAION-C. This dataset consists of over 336,000 images uniformly sampled across the 16 superclasses defined for LAION-C.

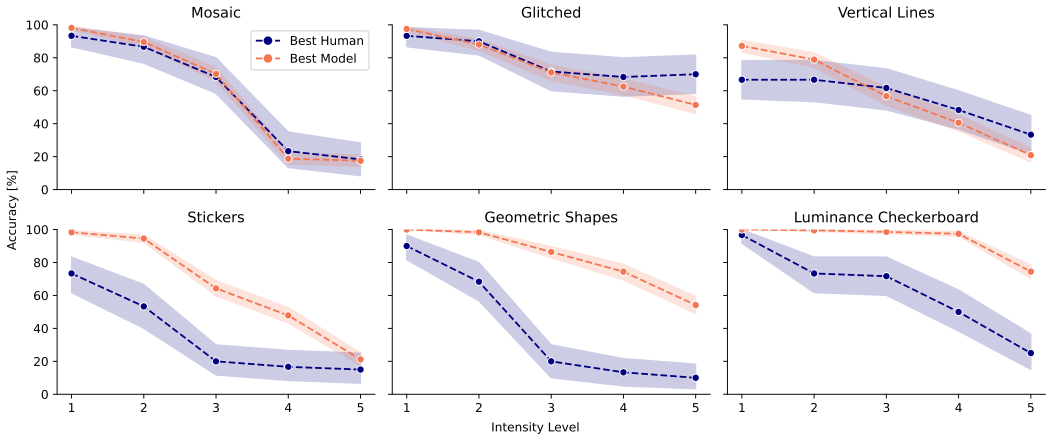

Figure 5: **Human vs. machine accuracy on all distortions.** For each LAION-C distortion, we plot the distortion intensity against the accuracy of the best human and the best model in this condition. The shaded regions indicate the 95% confidence intervals around the means. On the Mosaic, Glitched and Vertical Lines distortions, humans and machines perform similarly, whereas the best model vastly outperforms the best human observer on the Stickers, Geometric Shapes, and Luminance Checkerboard distortions.

As shown in Tab. 2, fine-tuning the model results in substantial accuracy gains, which define an upper bound on LAION-C accuracy that no normal model can be expected to achieve. Notably, these accuracy gains are particularly pronounced on higher-intensity distortions, as detailed in Tab. 4, where accuracy is broken down by distortion intensity. The fine-tuned model likely achieves such good performance by employing un-human-like (or "spurious") features, but the purpose of this experiment is *not* to suggest that fine-tuning on LAION is a sensible approach to improve OOD robustness, but to quantify how much learnable signal is left. LAION-C provides meaningful robustness tests without being intractable, making it a valuable tool for a more comprehensive evaluation of model performance under difficult conditions.

3.5 HUMAN AND MACHINE VISION DISCREPANCY

Accuracy Differences. In Fig. 5, we summarize how our suite of models performs in terms of classification accuracy, compared to the human participants in our psychophysical experiment. We report the best performances, since averages would be unfairly influenced by some older models we included as points of comparison. In Fig. 10, we provide a more detailed breakdown of performance by model. While human observers still outperform most vision models on images with Mosaic or Glitched distortions, the best models match (or even slightly surpass) human performance. For distortions involving occlusion and luminance manipulations, the vision models typically achieve higher accuracy than humans. Overall, current state-of-the-art vision models now match or even outperform human observers in OOD scenarios, including on our synthetic distortions, which they likely have never encountered during training—a stark contrast to just a few years ago, when humans were still vastly outperforming models (Geirhos et al., 2018; Dodge & Karam, 2019; Taori et al., 2020; Jang & Tong, 2024).

Occlusion and Luminance Manipulations. For distortions involving occlusions, such as Stickers and Geometric Shapes, models usually match or exceed human performance (see second row of Fig. 5). One possible hypothesis is that this can be attributed to the robustness that models develop after e.g., masked image modeling (MIM) (Fang et al., 2023; 2024). The fact that models perform so much better than humans on partially occluded images implies that models use different features than humans. For example, for the Stickers distortion, certain ViT models outperform humans, likely due to their ability to focus on those parts of the image background that remain visible despite the occlusions. As shown in Fig. 1, the stickers occlude nearly the entire image on higher intensity levels, and little to no meaningful object information is retained. Nevertheless, certain models are still able to correctly classify the image based on subtle background cues. This indicates that while mod-

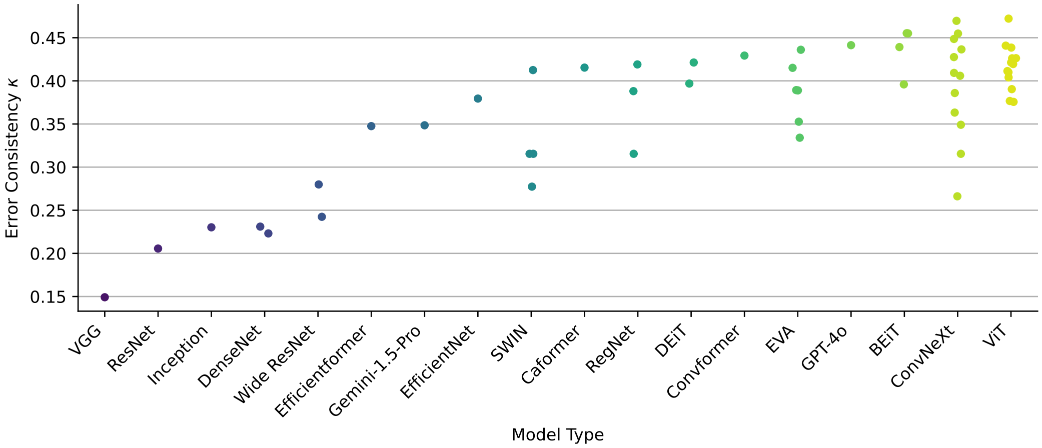

Figure 6: **Humans and models make different mistakes.** We analyze the agreement of error patterns between different families of vision models (see Tab. 6 for a complete list) and human observers. The error consistency (κ) could theoretically achieve a maximum value of 1, but in line with earlier work (Geirhos et al., 2021), the EC values range between 0 and 0.4, indicating that behavioral differences between humans and machines are still quite large. Marker colors encode model families.

els are performing well, they may be doing so by leveraging unintended shortcuts (Geirhos et al., 2020a), such as exploiting background information, when faced with severely occluded images. For the Luminance Checkerboard distortion, we observe that models from the ViT and EVA families outperform humans by a large margin. This advantage could potentially stem from their architectural features, such as self-attention mechanisms and patch-based processing (Fang et al., 2023; Dosovitskiy et al., 2021), which enable them to extract meaningful information from both light and dark regions independently, as well as handle subtle luminance variations. These capabilities give them a clear edge over humans and older models.

Performance on Complex Distortions. When analyzing more complex distortions such as Mosaic, Vertical Lines, and Glitched images (first row of Fig. 5), we observe that human performance is generally on par with the best-performing models. Especially at greater intensity levels, humans perform competitively, e.g., outperforming all models for the strongest Vertical Lines distortions. As we show in Fig. 10, the gap between humans and older models like the ResNet variants is particularly large on these complex distortions. However, modern model classes demonstrate substantial progress, approaching human-level performance even at higher intensity levels. While some margin for improvement remains, the narrowing gap suggests that achieving human-level robustness on classification tasks is no longer the primary challenge for state-of-the-art architectures.

Human-Machine Error Consistency on LAION-C. For a more fine-grained analysis of the behavioral agreement between models and human observers, we calculate error consistency as described in Sec. 2. As illustrated in Fig. 6, there is a high degree of variability in error consistency between human observers and different vision models ranging from 0 to 0.4. This indicates that while model families such as ViT and EVA rival or surpass human performance, they are approaching the task utilizing different strategies than humans, demonstrating less human-like behaviors. The observed value range matches the one found in previous work for older models and different image data (Geirhos et al., 2021). This again suggests that while recent developments have boosted model performance, these models have not become more human-like, as they follow alternative strategies.

4 DISCUSSION

Summary. Given that existing OOD benchmarks are often no longer OOD for models trained on web-scale datasets like LAION since distortions like blur and digital corruptions are commonplace on the web, we here introduce LAION-C. LAION-C is a benchmark designed to evaluate the robustness and generalization capabilities of modern vision models trained on web-scale datasets. Our empirical results demonstrate that LAION-C is more challenging for a representative suite of vision models than the existing ImageNet-C benchmark, particularly at higher distortion intensity levels. Our thorough human evaluation in a highly controlled psychophysical laboratory totaling 11,400 trials shows that *the best models often outperform even the best human observers*. While they do not always follow similar strategies (as indicated by error consistency analysis), this reassuring finding indicates that the best models have indeed substantially progressed in their ability to handle unexpected input and are not just getting better on in-distribution distortions. Given that the LAION-C benchmark dataset, by virtue of its construction, serves as a better proxy for a model's ability to recognize objects despite an unexpected distortion, we recommend it as an OOD benchmark for current and future web-scale vision models.

Limitations. While we have shown that LAION-C can effectively reveal shortcomings in model robustness, we have not yet fully explored why certain models underperform on specific distortions. Although our empirical results provide valuable insights, further investigation is required to clarify which visual cues the models rely on under different conditions. Such an analysis could inform the development of new inductive biases or architectural improvements, since a better understanding of these mechanisms could lead to improvements in both model interpretability and robustness. Given our current focus on introducing the dataset, this was not fully addressed here, but could be an area for future exploration. Furthermore, it is an open question what the performance limit on LAION-C looks like. Since fine-tuning models on LAION-C results in significant performance gains, particularly at higher distortion levels, there clearly is potential for optimization through advanced training techniques. However, how to further improve generalization across OOD scenarios, especially to enhance the models' ability to handle the novel distortions presented by LAION-C, remains an open question that warrants further exploration. To retain its value as an OOD benchmark, LAION-C should not be used as a training or fine-tuning dataset (except for analysis purposes).

Conclusion and outlook. Just a few years ago, early investigations into generalization abilities of deep neural networks showed humans vastly outperforming the best models (Geirhos et al., 2018; Dodge & Karam, 2019). Fast-forwarding to today, LAION-C shows that the best models either match or outperform human performance on challenging OOD distortions. This finding is reassuring in the light of growing concerns about the quality of existing evaluation datasets, including the concern that OOD datasets like ImageNet-C may no longer serve their original purpose in the era of web-scale training datasets. Our findings indicate that the often *super-human performance* of modern models is achieved through *super-human strategies*: Models use a variety of image cues—including, in all likelihood, background pixels to perform well on some of our challenging datasets. Given their high performance across the board, they no longer rely on a single strategy that fails when faced with a test case that distorts a particular image cue. This marks a paradigm shift in OOD generalization: From humans outperforming models to models outperforming humans, from relying on a single strategy to a diverse set of multiple robust strategies, and from OOD benchmarking measuring progress towards human-like object recognition to better performance now indicating super-human (in other words, *less human-like*) vision models.

CODE AND DATASET AVAILABILITY

We will publicly release the dataset and the code to generate distortions and evaluate models after acceptance of the paper. During the anonymous review period, the code is available from the supplementary material.

REFERENCES

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers. In *ICLR*, 2022.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint*, 2020. URL https://arxiv.org/abs/2006.07159.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 1960.

Samuel Dodge and Lina Karam. Human and dnn classification performance on images with quality distortions: A comparative study. *ACM Transactions on Applied Perception (TAP)*, 2019.

Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 2024.

Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological cybernetics*, 1975.

Robert Geirhos, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In *NeurIPS*, 2018.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020a.

Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In *NeurIPS*, 2020b.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In *NeurIPS*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR (Poster)*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021b.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint*, 2017. URL `https://arxiv.org/abs/1704.04861`.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

Hojin Jang and Frank Tong. Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks. *Nature Communications*, 15(1):1989, 2024.

Max Kaufmann, Daniel Kang, Yi Sun, Steven Basart, Xuwang Yin, Mantas Mazeika, Akul Arora, Adam Dziedzic, Franziska Boenisch, Tom Brown, Jacob Steinhardt, and Dan Hendrycks. Testing robustness against unforeseen adversaries. *arXiv preprint*, 2019. URL `https://arxiv.org/abs/1908.08016`.

Mario Kleiner, David Brainard, Denis Pelli, Allen Ingling, Richard Murray, and Christopher Broussard. What's new in psychtoolbox-3. *Perception*, 2007.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. *arXiv preprint*, 2022. URL `https://arxiv.org/abs/2203.06026`.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022.

Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel. Does clip's generalization performance mainly stem from high train-test similarity? *arXiv preprint*, 2023. URL `https://arxiv.org/abs/2310.09562`.

Prasanna Mayilvahanan, Roland S. Zimmermann, Thaddäus Wiedemer, Evgenia Rusak, Attila Juhos, Matthias Bethge, and Wieland Brendel. In search of forgotten domain generalization. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. URL `https://openreview.net/forum?id=Bc2p8T4V32`.

Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. *Advances in Neural Information Processing Systems*, 34:3571–3583, 2021.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. In *ICLR*, 2023.

OpenAI. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`, 2024. Accessed: 2024-11-19.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *CVPR*, 2022.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint*, 2022. URL `https://arxiv.org/abs/2208.06366`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 2015.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 2018.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022.

Chase T. Glitch this, 2020. URL `https://github.com/TotallyNotChase/glitch-this`. MIT License.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *NeurIPS*, 2020.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Felix A Wichmann and Robert Geirhos. Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, 2023.

Ross Wightman. The timm (pytorch image models) leaderboard. `https://huggingface.co/spaces/timm/leaderboard`, 2024.

Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint*, 2021. URL `https://arxiv.org/abs/2110.00476`.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 2014.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint*, 2016. URL `https://arxiv.org/abs/1605.07146`.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *CVPR*, 2022.

A APPENDIX

A.1 RELATED WORK

OOD generalization ability of vision models. As deep learning has advanced to the point where models can reliably generalize to data that matches their training distribution or even exceed the quality of the original labels (Beyer et al., 2020), OOD-robustness, as measured by specifically designed OOD test sets, has moved to the center stage of computer vision research. In particular, ImageNet-C (Hendrycks & Dietterich, 2019), a dataset containing images from the test-set of ImageNet to which various fairly natural corruptions such as blurring and pixelation were applied, has long been the gold standard for assessing OOD-performance, to the point where data augmentations proposed to increase OOD robustness were found to only work well because they are similar to the ImageNet-C corruptions (Mintun et al., 2021). In contrast, ImageNet-R (Hendrycks et al., 2021a) instead provides a more complex distribution shift by collecting different renditions of the target classes such as sculptures and paintings, instead of photos. A more subtle distribution shift which still caused considerable drops in model performance for ImageNet-trained models, was proposed by Recht et al. (2019). They collected ImageNetV2, a new test set for ImageNet that should theoretically not differ from the ImageNet test set at all, because it was collected with a very similar methodology, but revealed that models do perform slightly worse on ImageNetV2 than on the original test set. Hendrycks et al. (2021b) proposed two other OOD-test sets which do not rely on synthetic image manipulations but instead consist of natural images that are in some sense OOD relative to ImageNet, either by virtue of displaying object classes not present in ImageNet (ImageNet-O) or by showing an object of an ImageNet-class in a scene that is weird enough to fool most models (ImageNet-A). What all of these datasets have in common is that, by design, they provide distribution shifts *relative to ImageNet*. But with the rapid evolution of deep learning, these traditional benchmarks have become less challenging for state-of-the-art vision models trained on expansive web-scale datasets (Radford et al., 2021). While it is to some degree possible to predict a model's OOD generalization from its training set performance (Taori et al., 2020), the only reliable measurements of this capability stem from empirical evaluations of models on OOD datasets. Our work addresses this need by introducing LAION-C, a dataset that incorporates novel and complex synthetic distortions tailored to challenge even advanced vision systems.

Advancement in visual foundation models The rise of visual foundation models such as Vision Transformers (ViT) (Dosovitskiy et al., 2021), ConvNeXt (Liu et al., 2022) and BeiT (Bao et al., 2022) has redefined what constitutes standard performance across many visual tasks. These improvements in performance partially stem from architectural innovations and parameter optimization, but were mostly powered by the effective leveraging of unprecedented dataset sizes (Zhai et al., 2022). However, because visual foundation models were trained on web-scale datasets, the extent of their generalization capability remains underexplored.

Comparing human vs. machine perception. Deep Neural Networks were originally conceived as models of human vision (Fukushima, 1975) and were found to be the best available models for neuronal activity in the primate visual cortex (Yamins et al., 2014), even if not trained for this task. Today, there is a growing body of research dedicated to evaluating the adequacy of neural networks as behavioral models of human core object recognition (Doerig et al., 2023; Schrimpf et al., 2018; Wichmann & Geirhos, 2023; Muttenthaler et al., 2023). Building upon the findings of Geirhos et al. (2021), who illustrate the narrowing of the behavioral gap between humans and machines in terms of their error consistency, our study further explores this dynamic utilizing LAION-C. We conducted a systematic analysis of differences in perception between human and machine observers, and assessed if the behavioral gap is closing further, as well as highlighting the persistent cognitive differences between humans and machines.

A.2 EXPERIMENT PROCEDURE AND PARTICIPANT INCENTIVES

Participant recruitment and setup. We recruited 20 participants (10 female) from the university student body via mailing lists. All participants were screened to ensure normal or corrected vision and no color blindness, and gave informed consent to participate. One participant was excluded post-hoc due to reporting extreme tiredness. Our experiments were conducted in a darkened cabin,

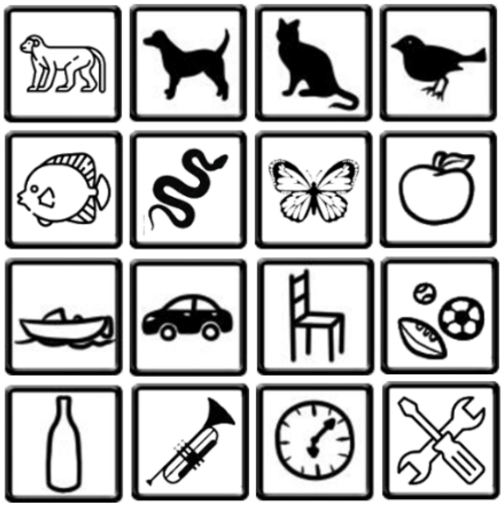

Figure 7: **Interface presented to participants.** This figure illustrates the icon layout as displayed to participants during the study. The grid is adapted from (Geirhos et al., 2018), while most of the categories and therefore symbols are different.

using a 22” VIEWPixx 3D light LCD monitor (VPixx Technologies, Saint-Bruno, Canada) at a refresh rate of 120 Hz (scanning backlight mode on). The screen measures 484×302 mm, at a resolution of 1920×1200 pixels. Stimuli were presented foveally in the center of the screen, with a viewing distance of 65 cm, resulting in 5 ° of visual angle. In line with earlier experiments, the background was set to a grey value of 0.454 in the $[0, 1]$ range. A chin rest was used to maintain a fixed viewing distance and angle. The experiment was implemented using the Psychophysics Toolbox (Kleiner et al., 2007, version 3.0.12) in MATLAB (Release 2016a, The MathWorks, Inc., Natick, Massachusetts, United States) using a 12-core desktop computer (AMD HD7970 graphics card “Tahiti” by AMD, Sunnyvale, California, United States) running Kubuntu 14.04 LTS.

The entire classification task, including both the warm-up and main experiment phases, was organized into 12 blocks. In each block, participants were shown images from the 16 superclasses for 2.5 seconds—a duration empirically determined to balance efficient overall experiment length with sufficient exposure time allowing for viewing and consideration time. After each image, the 16 corresponding class icons were displayed on screen, allowing participants 2 seconds to classify each image into one of these categories. The icons were organized in a layout that roughly grouped them by size and general category (e.g., animals and objects), as illustrated in image Fig. 7. To encourage responses rather than leaving selections blank, a message was displayed at the top of the screen 0.75 second before icon display time ended, prompting participants to make a choice. At the end of each block, if a participant surpassed the 90% accuracy threshold calibrated using internal baseline performance data, they received an encouraging on-screen message (“Congratulations! You just earned some extra money!”) along with a $0.50 bonus per block to incentivize higher performance.

Warm-up session and main experiment. The experiment began with a 10-minute warm-up session, allowing participants to familiarize themselves with the icon layouts and the classification task procedure through two practice blocks. Each practice block contained 45 images, with one block consisting of clean images and the other of augmented images. To avoid test-time adaptation, the augmentations used during the warm-up phase differed from those in the actual trials. The images used for the practice trials were also a subsample of the ImageNet validation dataset, but did not overlap with those used in the main experiment.

Following the warm-up, the main experiment proceeded consisting of 10 blocks, each block comprising 60 images. Each set of 5 blocks was augmented using a consistent method, with a balanced distribution across different intensity levels and superclasses. To avoid fatigue, participants were allowed an unlimited amount of time to rest between blocks and encouraged to rest their eyes or accomodate elsewhere.

A.3 ERROR CONSISTENCY

Here, we provide a more detailed explanation of error consistency (EC), summarizing Geirhos et al. (2020b). The EC between two observers which both classified a sequence of samples is obtained by first using the necessary ground-truth labels to assess which images each observer classified correctly. A trial increases the agreement if both observers solved it correctly, or if they both failed (and decreases it if only one of them got the trial correct while the other one failed). One then calculates how much more agreement was observed between the two observers relative to the agreement expected by chance. This is done by calculating Cohen's Kappa (Cohen, 1960), which is defined as $\kappa = \frac{p_o - p_e}{1 - p_e}$, where p_o is the (empirically measured) proportion of agreement-trials and p_e is the (theoretical) expected agreement when modeling both observers as independent binomial observers. κ takes on values between -1 and 1, with a higher κ signifying higher levels of agreement, and a κ of 0 indicating that a pair of observers does not agree more frequently than one would expect by chance, given their marginal correctness probabilities.

In this work, we calculate the error consistency between model responses and human classification data. To do this, we first collect all human responses. Since each human participant saw a fresh set of stimuli, we thus obtain exactly one human response per image. We then calculate each model's EC to this list of human responses.

A.4 AUGMENTATION DESIGNS

- **Mosaic:** The image is divided into an $n \times n$ grid, where each tile is replaced by a patch from the ImageNet validation set whose average color best matches that of the tile. The values of n per intensity level are:
    - Level 1: $n = 4$
    - Level 2: $n = 6$
    - Level 3: $n = 8$
    - Level 4: $n = 16$
    - Level 5: $n = 28$
- **Glitched:** Alternating rows are replaced with black pixels to create a scan line effect. Pixel shifts and color channel offsets are applied to random regions as follows:
    - Level 1: Shift up to 8% of image width in 4 regions, ±4 pixel channel offset.
    - Level 2: Shift up to 32% of image width in 8 regions, ±8 pixel channel offset.
    - Level 3: Shift up to 50% of image width in 10 regions, ±10 pixel channel offset.
    - Level 4: Shift up to 128% of image width in 16 regions, ±16 pixel channel offset.
    - Level 5: Shift up to 200% of image width in 20 regions, ±20 pixel channel offset.

  The implementation is inspired by T (2020)
- **Vertical Lines:** The image is divided into vertical sections, each of which is further subdivided into smaller sections along the y-axis (called y-steps). A vertical line is drawn within each y-step with a slight x-offset based on the intensity level. The line color is determined by the average color of that section. The parameters for each intensity level are:
    - Level 1: 224 vertical sections, with 1-pixel steps along the y-axis.
    - Level 2: 178 vertical sections, with 2-pixel steps along the y-axis.
    - Level 3: 112 vertical sections, with 4-pixel steps along the y-axis.
    - Level 4: 84 vertical sections, with 6-pixel steps along the y-axis.
    - Level 5: 60 vertical sections, with 8-pixel steps along the y-axis.
- **Luminance Checkerboard:** The image is divided into a 14×14 grid, and the luminance of each tile is altered in a checkerboard pattern. The luminance variation per intensity level is:
    - Level 1: ±50.
    - Level 2: ±50–100.
    - Level 3: ±100–125.
    - Level 4: ±125–150.
    - Level 5: ±150–255.

Table 3: **Occlusion ratio of objects in Stickers and Geometric Shapes distortions.** We calculated the object occlusion ratio for the Stickers and Geometric Shapes corruptions as an additional quantitative measurement of the distortion strength.

Intensity Level	Geometric Shapes (%)	Stickers (%)
1	61.88	65.83
2	72.51	76.52
3	85.35	86.19
4	90.16	89.54
5	93.21	91.63

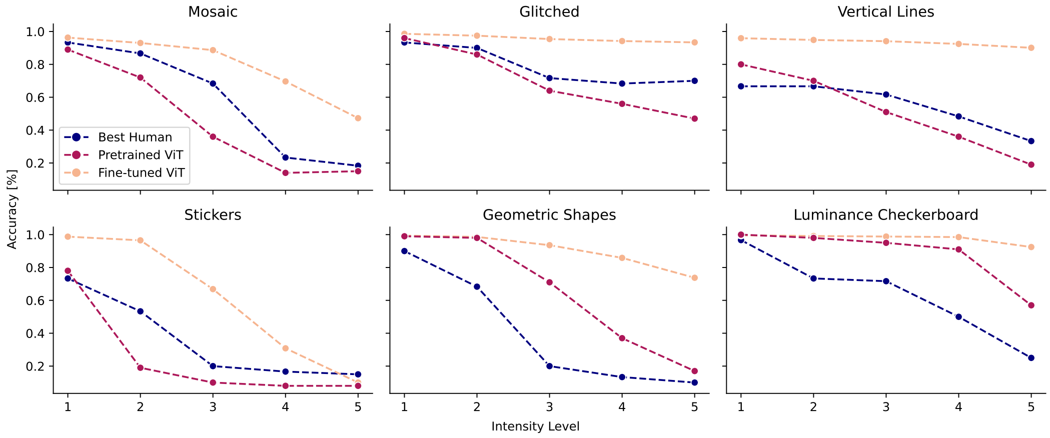

Figure 8: **LAION-C can be solved.** For every distortion, we plot the accuracy of our reference model (ViT-H-P14-336-CLIP-LAION-IN12K) before and after fine-tuning, in comparison to the best human participant for reference. Most distortions can be learned perfectly, only the Stickers and Mosaic distortions might have been too difficult at the highest intensity levels. Further performance gains might be possible with more careful fine-tuning.

- **Stickers:** 16×16 pixel image patches from the ImageNet validation set are randomly placed with uniform probability across the image. The number of patches per intensity level is:
  - Level 1: 100 patches.
  - Level 2: 200 patches.
  - Level 3: 400 patches.
  - Level 4: 600 patches.
  - Level 5: 1200 patches.

  For an estimate of the occlusion ration of the objects per intensity level, see Tab. 3.

- **Geometric Shapes:** Random geometric shapes (triangle, square, star, circle) of varied colors and sizes are introduced. The number of shapes per intensity level is:
  - Level 1: 150 shapes.
  - Level 2: 300 shapes.
  - Level 3: 600 shapes.
  - Level 4: 800 shapes.
  - Level 5: 1000 shapes.

  For an estimate of the occlusion ration of the objects per intensity level, see Tab. 3.

Table 4: **Model (ViT) Accuracy Before and After Fine-Tuning on LAION-C.** The high accuracies after fine-tuning indicate that even though the dataset is challenging, there is, in principle, enough signal left to perform well on LAION-C.

	Intensity Level	Accuracy Before (%)	Accuracy After (%)
Mosaic	1	89.0	96.3
	2	71.9	93.0
	3	35.8	88.7
	4	14.3	69.6
	5	14.7	47.7
Vertical Lines	1	79.9	95.9
	2	70.1	94.9
	3	50.8	94.1
	4	36.1	92.4
	5	19.4	90.0
Glitched	1	95.9	98.6
	2	86.2	97.5
	3	63.6	95.4
	4	55.6	94.2
	5	47.1	93.4
Luminance Checkerboard	1	99.7	99.6
	2	98.4	99.2
	3	95.1	98.8
	4	90.7	98.5
	5	56.6	92.5
Geometric Shapes	1	30.9	99.4
	2	11.2	98.6
	3	6.7	93.6
	4	6.6	85.9
	5	6.3	73.7
Sticker	1	97.3	98.8
	2	77.8	96.5
	3	28.7	63.7
	4	14.9	31.8
	5	8.1	14.3

A.5 ACCURACY

To demonstrate the value of LAION-C as a benchmark for evaluating model robustness, we analyze how model performance on LAION-C correlates with that on ImageNet-C. Grounding our comparison in models that have demonstrated a baseline level of robustness on well-established benchmarks, we apply a threshold to include 40 models that achieved at least 60% accuracy on ImageNet-C.
As shown in Fig. 9, the majority of data points lie above the identity line representing performance alignment on LAION-C and ImageNet-C. The gradual slope of the data points, combined with their positioning, indicates that models generally perform better on ImageNet-C, while their performance on LAION-C is more dispersed and often substantially lower.
This broader distribution of performance highlights that LAION-C introduces more challenging distortions, prompting models to exhibit greater variability in robustness. The moderate Kendall's tau coefficient (τ = 0.66) between the models' performances on LAION-C and ImageNet-C further underscores this, indicating notable pairwise differences in how models rank across these two benchmarks, unearthing vulnerabilities that are less pronounced on ImageNet-C. These results demonstrate the necessity of LAION-C as a complementary benchmark for a more comprehensive evaluation of model robustness.

A.6 BREAKDOWN OF MODEL PERFORMANCE

Evaluating VLMs To evaluate GPT-4o (OpenAI, 2024) and Gemini 1.5 Pro (Team et al., 2024) on LAION-C, we decided to test a random subsample of the full dataset, consisting of 100 images per category, which were then tested on all corruptions and intensity levels, resulting in a total of

48,000 images. For ImageNet-C, we limited ourselves to only 10 images per class, to get an initial ballpark estimate of performance.

We employed the following system prompt, in line with our human experiments, during which participants were also shown examples:

System Prompt:
You are an image-recognition API.
You are always asked to classify the main object of images into one of 16 mutually exclusive categories.
In some images, the distortion may be so strong that you might not recognize anything. If you're unsure, provide your best guess - you always have to pick exactly one of the 16 categories.
The 16 categories are: primate, dog, cat, bird, fish, snake, butterfly, fruit, boat, vehicle, chair, ball, bottle, instrument, timekeeper, tool.
Here is a list of characterizations of every such category:
primate: a primate, like e.g. monkeys, chimpanzees, Orang-Utans etc.
dog: a dog, like e.g. german shepherd, labrador, golden retriever etc.
cat: a cat, like e.g. domestic cat, lion, cheetah etc.
bird: a bird, like e.g. songbird, eagle, chicken etc.
fish: a fish, like e.g. trout, shark, whale etc.
snake: a snake, like e.g. viper, cobra, seasnake etc.
butterfly: a butterfly, like e.g. monarch, cabbage butterfly, ringlet etc.
fruit: a fruit, like e.g. apple, orange, pineapple etc.
boat: a boat, like e.g. ship, gondola, fireboat etc.
vehicle: a vehicle, like e.g. truck, van, sports car etc.
chair: a chair, like e.g. bench, throne, couch etc.
ball: a ball (or a person playing with a ball), like e.g. soccer ball, football, tennis ball etc.
bottle: a bottle, like e.g. water bottle, jug, pill bottle etc.
instrument: a musical instrument (or a person playing an instrument), like e.g. sax, flute, harp etc.
timekeeper: a timekeeper, like e.g. clock, watch, sundial etc.
tool: a tool (or a person using a tool), like e.g. hammer, power drill, chainsaw etc.
Since you are an API, you always respond with minimal messages that contain exactly one word, which is the category name.

User Prompt:
What is the main object in this image? Categories are: primate, dog, cat, bird, fish, snake, butterfly, fruit, boat, vehicle, chair, ball, bottle, instrument, timekeeper, tool.

Table 5: **Model performance on LAION-C correlates with other OOD benchmarks.** We evaluated a suite of 18 models (ViT and ConvNeXt variants trained on either LAION-2B or ImageNet) on IN-C, IN-A, IN-R, IN-Sketch and IN-Val. Evidently, the correlations between all of these OOD benchmarks are high, indicating that they measure related quantities.

	IN-C	LAION-C	IN-A	IN-R	IN-Sketch	IN-val
IN-C	1.00	0.86	0.88	0.91	0.86	0.90
LAION-C	0.86	1.00	0.69	0.70	0.81	0.72
IN-A	0.88	0.69	1.00	0.99	0.94	1.00
IN-R	0.91	0.70	0.99	1.00	0.93	0.99
IN-Sketch	0.86	0.81	0.94	0.93	1.00	0.95
IN-val	0.90	0.72	1.00	0.99	0.95	1.00

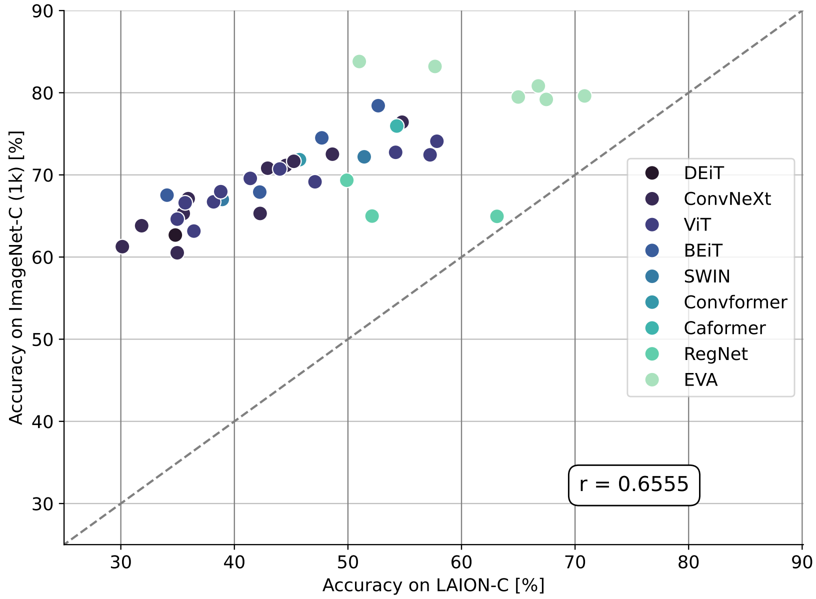

Figure 9: **Performance Divergence of Models on LAION-C and ImageNet-C (1k classes).** The figure illustrates the scattered performance of models across the ImageNet-C and LAION-C dataset, where a Kendall's tau coefficient (τ) of 0.66 and the shallow slope indicate a dispersed performance on LAION-C. To provide a clearer trend and to better visualize the dispersion, we supplement the suite of models with additional top-performing models sourced from the timm leaderboard (Wightman, 2024), bringing the total number of models to 40 (see Tab. 6 for a complete list).

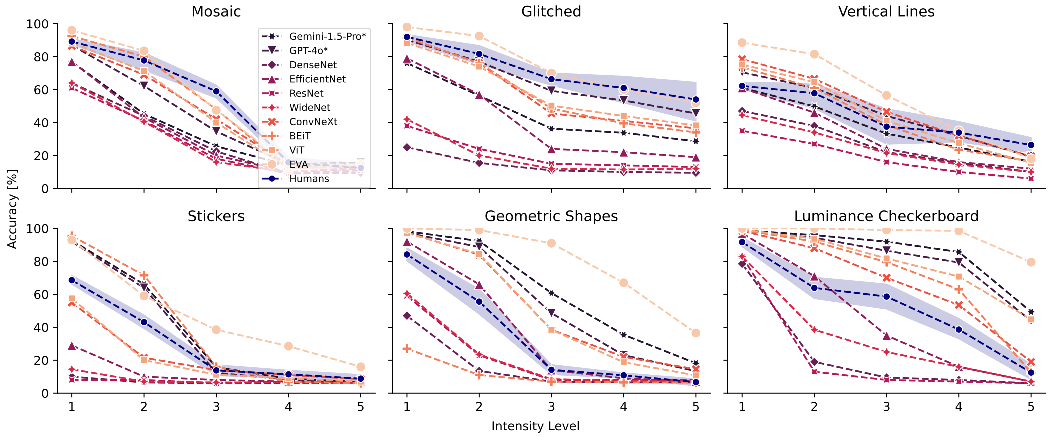

Figure 10: **Model performance on LAION-C.** Analogous to Fig. 5, we relate distortion intensity level to classification accuracy for the different distortions, showing the different models individually. The shaded region around human performance corresponds to the 95% confidence interval, which we omit for the models for better visibility.

A.7 MODELS

Table 6: **Model overview.** For each model used in our evaluation, we show the full model names, as used in timm, an abbreviated name used in the main text and a description of the model. While the first 16 models were used in all analyses and figures, the rest was only used for selective analyses such as Fig. 6.

Abbreviation	Full Model Name	Description
EVA-G-P14-560-M30M-IN22K	eva_giant_patch14_560.m30m_ft_in22k_in1k	EVA giant model, patch size 14, pre-trained with masked image modeling (MIM) on a Merged-30M dataset, fine-tuned on ImageNet-22k and ImageNet-1k (Fang et al., 2023).
EVA02-L-P14-448-MIM-M38M-IN22K	eva02_large_patch14_448.mim_m38m_ft_in22k_in1k	EVA02 large model, patch size 14, pre-trained with masked image modeling (MIM) on a Merged-38M dataset, fine-tuned on ImageNet-22k and ImageNet-1k (Fang et al., 2024).
VIT-H-P14-336-CLIP-LAION-IN12K	vit_huge_patch14_clip_336.laion2b_ft_in12k_in1k	Vision Transformer (VIT) huge model, patch size 14, pre-trained on LAION-2B dataset using OpenCLIP, fine-tuned on ImageNet-12k and ImageNet-1k (Dosovitskiy et al., 2021).
VIT-L-P14-224-CLIP-OPENAI-IN12K	vit_large_patch14_clip_224.openai_ft_in12k_in1k	Vision Transformer large model, patch size 14, pre-trained on WIT-400M using CLIP, fine-tuned on ImageNet-12k and ImageNet-1k (Dosovitskiy et al., 2021).
VIT-B-P32-384-CLIP-LAION-IN12K	vit_base_patch32_clip_384.laion2b_ft_in12k_in1k	Vision Transformer base model, patch size 32, pretrained on LAION-2B using OpenCLIP,fine-tuned on ImageNet-12k and ImageNet-1k (Dosovitskiy et al., 2021).
VIT-B-P16-224-AUGREG-IN21K	vit_base_patch16_224.augreg2_in21k_ft_in1k	Vision Transformer base model, patch size 16, trained on ImageNet-21k and fine tuned on ImageNet-1k (Dosovitskiy et al., 2021).
BEITV2-L-P16-224-IN1K	beitv2_large_patch16_224.in1k_ft_in1k	BEiTv2 large model, patch size 16, trained on ImageNet-1k, fine-tuned on ImageNet-22k and ImageNet-1k (Bao et al., 2022; Peng et al., 2022).
BEITV2-B-P16-224-IN1K	beitv2_base_patch16_224.in1k_ft_in1k	BEiTv2 base model, patch size 16, trained on ImageNet-1k, fine-tuned on ImageNet-22k and ImageNet-1k (Bao et al., 2022; Peng et al., 2022).
CONV-XXL-CLIP-LAION-IN1K	convnext_xxlarge.clip_laion2b_soup_ft_in1k	ConvNeXt xxlarge model, pre-trained using OpenCLIP on LAION-2B, fine-tuned on ImageNet-1k (Liu et al., 2022).
CONV-B-CLIP-LAION-AUGREG-IN12K	convnext_base.clip_laion2b_augreg_ft_in12k_in1k_384	ConvNeXt base model,pre-trained using OpenCLIP on LAION-2B, fine-tuned on ImageNet-12k and ImageNet-1k (Liu et al., 2022).
WRN101-2-TV-IN1K	wide_resnet101_2.tv_in1k	Wide ResNet-101 model, trained on ImageNet-1k, with original torchvision model weight (He et al., 2016; Zagoruyko & Komodakis, 2016).
WRN50-2-RACM-IN1K	wide_resnet50_2.racm_in1k	Wide ResNet-50 model, trained with RandAugment RACM recipe on ImageNet-1k (He et al., 2016; Zagoruyko & Komodakis, 2016).
RN50-A1-IN1K	resnet50.a1_in1k	ResNet-50 model trained on ImageNet-1k (He et al., 2016; Wightman et al., 2021).
EFF-B3-RA2-IN1K	efficientnet_b3.ra2_in1k	EfficientNet-B3 model, trained with RandAugment RA2 recipe on ImageNet-1k (Tan & Le, 2019).
DN201-TV-IN1K	densenet201.tv_in1k	DenseNet-201, DenseNet pre-trained on ImageNet-1k (Huang et al., 2017).
DN161-TV-IN1K	densenet161.tv_in1k	DenseNet-161, DenseNet model pre-trained on ImageNet-1k (Huang et al., 2017).
GPT-4o	gpt-4o-2024-08-06	At the time of writing, the most recent snapshot of OpenAI's flagship model (OpenAI, 2024). Only evaluated on 48,000 LAION-C samples and 12,000 ImageNet-C samples.
Gemini-1.5-Pro	gemini-1.5-pro-002	At the time of writing, the most recent stable version of Google's Gemini model (Team et al., 2024). Only evaluated on 48,000 LAION-C samples and 12,000 ImageNet-C samples.
	convnextv2_pico.fcmae_ft_in1k	
	convnextv2_tiny.fcmae_ft_in22k_in1k	
	convnext_base.fb_in22k_ft_in1k	
	convnext_large_mlp.clip_laion2b_augreg_ft_in1k_384	
	convnext_large_mlp.clip_laion2b_soup_ft_in12k_in1k_384	
	convnext_tiny.in12k_ft_in1k	
	convnext_small.fb_in22k_ft_in1k_384	
	convnext_xlarge.fb_in22k_ft_in1k	
	convnext_small.in12k_ft_in1k_384	
	convnextv2_large.fcmae_ft_in22k_in1k_384	
	vit_betwixt_patch16_reg4_gap_256.sbb2_e200_in12k_ft_in1k	
	vit_mediumd_patch16_rope_reg1_gap_256.sbb_in1k	
	vit_wee_patch16_reg1_gap_256.sbb_in1k	
	vit_mediumd_patch16_reg4_gap_256.sbb2_e200_in12k_ft_in1k	
	vit_mediumd_patch16_reg4_gap_256.sbb_in12k	
	vit_pwee_patch16_reg1_gap_256.sbb_in1k	
	vit_betwixt_patch16_rope_reg4_gap_256.sbb_in1k	
	vit_betwixt_patch16_reg4_gap_256.sbb_in12k_ft_in1k	
	maxxvitv2_rmlp_base_rw_384.sw_in12k_ft_in1k	
	vgg19_bn.tv_in1k	
	regnety_1280.swag_lc_in1k	
	regnety_1280.swag_ft_in1k	
	regnety_320.swag_ft_in1k	
	inception_v3.tf_adv_in1k	
	beit_base_patch16_224.in22k_ft_in22k_in1k	
	beit_large_patch16_512.in22k_ft_in22k_in1k	
	deit3_large_patch16_384.fb_in22k_ft_in1k	
	deit_base_distilled_patch16_224.fb_in1k	
	swin_base_patch4_window7_224.ms_in22k_ft_in1k	
	swinv2_base_window12to24_192to384.ms_in22k_ft_in1k	
	swinv2_large_window12to24_192to384.ms_in22k_ft_in1k	
	eva_large_patch14_336.in22k_ft_in1k	
	convformer_b36.sail_in22k_ft_in1k_384	
	caformer_b36.sail_in22k_ft_in1k_384	
	efficientformerv2_s2.snap_dist_in1k	